ORIGINAL RESEARCH ARTICLE

# Bayesian modeling of factorial time-course data with applications to a bone aging gene expression study

Joseph Wu[a,b], Mayetri Gupta[c] , Amira I. Hussein[d] and Louis Gerstenfeld[d]

[a]Boston University School of Public Health, Boston, MA, U. S. A.; [b]Pfizer, Inc., Groton, CT, U.S.A; [c]University of Glasgow, Glasgow, U. K.; [d]Boston University School of Medicine, Boston, MA, U. S. A.

**ARTICLE HISTORY**

Compiled May 19, 2020

**ABSTRACT**

Many scientific studies, especially in the biomedical sciences, generate data measured simultaneously over a multitude of units, over a period of time, and under different conditions or combinations of factors. Often, an important question of interest asked relates to which units behave similarly under different conditions, but measuring the variation over time complicates the analysis significantly. In this article we address such a problem arising from a gene expression study relating to bone aging, and develop a Bayesian statistical method that can simultaneously detect and uncover signals on three levels within such data: factorial, longitudinal, and transcriptional. Our model framework considers both cluster and time-point-specific parameters and these parameters uniquely determine the shapes of the temporal gene expression profiles, allowing the discovery and characterization of latent gene clusters based on similar underlying biological mechanisms. Our methodology was successfully applied to discover transcriptional networks in a microarray data set comparing the transcriptomic changes that occurred during bone aging in male and female mice expressing one or both copies of the bromodomain (Brd2) gene, a transcriptional regulator which exhibits an age-dependent sex-linked bone loss phenotype.

**KEYWORDS**

Factorial Designs; Markov chain Monte Carlo; Microarrays; Mixture models;

CONTACT Mayetri Gupta Email: mayetri.gupta@glasgow.ac.uk

Model-based Clustering; Temporal gene expression.

## 1. Introduction

Many modern scientific experiments aim to collect, and analyze, longitudinally measured data under different conditions (or combinations of factors), simultaneously over a large number of units. An important question of interest then relates to assessing whether there are groups of units that show similar patterns across time and over conditions. For instance, in experimental designs such as factorial time-course gene expression studies, scientists often want to assess how two factors interact in their time-specific effects on gene expression. At the same time, one may wish to identify genes that work together as part of the same transcriptional network.

An RNA microarray experiment provides a snapshot of gene expression in a cell, depicting how thousands of genes are simultaneously expressed or suppressed at a single point of time [22, 32]. Profiling dynamic transcriptional activity gives important insights into how genes respond over time to conditions such as exposure to pathogens, administration of a drug, disease progression, or even normal aging [17, 24]. Genes that are closely related in a regulatory network tend to behave similarly during expression or suppression. As a result, in a time-course microarray experiment, related genes tend to share similar temporal expression profiles [40]. With technological advances, studies of gene expression over time and under different combinations of factors are becoming economically feasible and researchers can choose from tools such as microarray analyses, which target the identification of known alleles, to newer approaches such as RNA sequencing (RNA-Seq), which provides a complete profile of a sample's transcriptome [16]. Regardless of the tool, gene expression studies generate large data sets for analysis, with multiple layers of complexity, ranging from design issues (e.g. multiple factor combinations), to high levels of correlatedness, due to temporal measurements of genes as well as their presence in the same biological pathway [30]. Powerful statistical approaches are then necessary to uncover complex mechanistic relationships between multiple genes.

In clinical settings, investigators may be interested in finding how genes are ex-

pressed differentially between experimental and control groups of subjects over the course of the experiment. For instance, investigators may want to know how differential gene expression profiles between an experimental group and a control group change with the addition of new experimental factors. The effect of an experimental treatment may differ between two age groups, sexes, or sets of genetic predisposition. One may want to detect possible interactions between an experimental treatment and another factor while concurrently grouping genes that show similar temporal expression profiles into clusters. Answers to these questions will have useful applications in the development of pharmaceutical compounds that exert targeted effects in one group of patients versus another. On the molecular level, this will help investigators unlock the biological mechanisms of a pharmaceutical compound among different subgroups of patients. Although the general framework and methods are proposed here in the context of gene expression microarrays, these can be adapted to sequencing-based experiments as well as in other applications where clustering of observations that are measured under different combinations of factors, and longitudinally, is needed. In the next section, we discuss some current statistical approaches to analyze temporal gene expression data before introducing our newly proposed model and methodology.

### *1.1. Background*

Methods have been developed in recent years to classify thousands of related genes exhibiting similar temporal profiles into clusters. Ramoni, Sebastiani and Kohane [25] assumed an auto-regressive model for the time series $AR(p)$ where the expression value at time $t$ was linearly related to the previous $p\,(t > p)$ expression values. Schliep, Schonhuth and Steinhoff [27], and Zeng and Garcia-Frias [37] developed a clustering algorithm using hidden Markov models and profile hidden Markov models to account for time-dependency. Clustering of genes under these models was based on similar profile dynamics. Another important class of methods applied non-parametric spline techniques to model the longitudinal expression profiles as continuous functions [20]. A number of methods incorporated transcription factor binding site motifs as regression covariates to find regulatory networks, such as through a Bayesian hidden component model [26] or a hidden Markov model for time-dependent gene expression data [11].

Ernst, Nau and Bar-Joseph [3] developed a deterministic data-driven approach, predefining a set of "gene profiles" to which genes were assigned by optimizing a distance-based criterion. Others have focused on modeling periodic gene expression data such as in the mitotic cell cycle [33] using first order auto-regressive models with random effects. More recent methods included jointly modeling longitudinal expression profiles with gene ontology (GO) tags from existing GO databases to improve clustering [5] and reconstructing time-course profiles into wavelet-based multi-resolution fractal features followed by mixture modeling [18]. A few methods were able to simultaneously infer differential expression and to account for gene clustering [35, 36]. Kayano et al. [15] proposed a functional logistic model based on elastic net regularization to identify genes with dynamic alterations between case and control subjects. Scholtens et al. [28] considered two binary factors in a longitudinal gene expression study but did not consider modeling temporal profiles and gene clustering by similar expression patterns to determine subgroups of genes from similar regulatory pathways. Zhou et al. [39] considered longitudinal gene expression as individual vectors and located an optimal multivariate ANOVA signal for each gene, subsequently classifying genes into five groups based on a series of non-parametric tests. These five groups referred to different resultant ANOVA models: interaction, main effects, or null models. However, their classification method was limited to testing these model signals rather than accounting for unknown temporal expression patterns that could reflect more diverse biological processes.

As experimental designs become more complex, as in a factorial time-course microarray experiment, a methodological gap still exists in determining functionally related gene clusters that may behave similarly over time and under different combinations of factors, through fully modeling the data structure. None of the approaches discussed above can be directly applied to such data to cluster genes into groups that share similar temporal patterns under specific factor combinations, without losing some information on the structure of the data. In this article, we propose a new Bayesian model-based approach to simultaneously estimate the longitudinal model signals under a factorial design and assign genes into biologically meaningful clusters. In this model, all information about gene expression is preserved and can be

interpreted at all three levels– longitudinal, factorial, and transcriptional. Although Bayesian mixture models have been the subject of much research in recent years [31], to our knowledge, such models have not yet been developed in the context of time-course data from factorial experiments. Lu and Huang [19] proposed a mixture of Bayesian mixed effects models to analyze viral load data. However, in their method, the total number of classes in the mixture needed to be fixed in advance, with the mean functions of each class being explicitly parametrized in the context of the application. To motivate our approach, we first discuss the structure of the microarray experiment and the data.

### *1.2. Bone aging microarray study in mice*

Our research was motivated from a collaboration with the Orthopaedic Research Laboratory at Boston University. Biologists were interested in comparing bone aging in male and female mice expressing one (mutant) or both (wild-type) copies of the bromodomain (Brd2) transcription factor. Mice with reduced expression of the transcriptional regulator Brd2 become obese, but show reduced hyper-inflammatory responses to stress and remain insulin sensitive [2, 34]. The primary objective of the experiment was to characterize the gene expression profiles over a time span in which peak bone mass was reached and skeletal growth was completed to the onset of early age-dependent changes in bone quality. Specifically, the study sought to differentiate the effects of sex and reduced expression of Brd2 on the global gene expression patterns obtained as the animals aged. Our approach aimed to identify clusters containing gene groups, which were associated with biological processes related to the development of bone-related diseases. Further details on the data are provided in Section 4.

The structure of the paper is outlined below. In Section 2, we describe the Bayesian factorial time-course mixture model for gene expression data and assess the performance of this model in simulation studies in Section 3. In Section 4, we apply this model to the mouse data, comparing the transcriptomic changes that occur during bone aging in wild-type and mutant versions of male and female mice. The ability of this modeling approach to identify the temporal relationships of the biological processes that are associated with sex and genotype is then assessed. Finally, in Section

5, we discuss the results, implications, and future directions of this approach.

## 2. Modelling framework and methodology

Our data set (Sec 1.2) is generated from a large-scale gene expression microarray experiment in mice under different conditions (the two factors being sex and mutation status) over time. The goal was to detect whether groups of genes behaved in a similar way over time, under a particular combination of factors, and differed from other sets of genes in their pattern of behavior. The transcript abundance was first quantified and standardized into gene expression values through a series of data processing steps (details given in Appendix B of the Supplementary Materials).

### *2.1. A factorial time-course mixture model*

Let $y_{gijtr}$ represent the log-normalized gene expression value for gene $g$ ($g = 1, ..., G$), measured at time point $t (t = 1, ..., T)$ given two factors, the levels of which are specified by $i$ and $j$, each taking values in the set $\{0, 1\}$. $r$ ($r = 1, ..., R$) denotes the replicate index. Theoretically, this setup could also be generalized to factors with more than 2 levels, as well as a larger number of factors. However, at present, we describe our methodology in the context of two factors for notational simplicity, and also because this is one of the most common types of factorial design.

For the present, let us assume that there are $K$ stable clusters of genes over the entire course of the experiment. Let $z_{gk} = 1$ (or 0) indicate that gene $g$ belongs (or does not belong) to cluster $k$ ($k = 1, ..., K$), with each gene belonging to exactly one of the clusters (this could be a singleton cluster), so that $\sum_{k=1}^{K} z_{gk} = 1$. With boldface letters representing vectors or matrices from now on, let $\boldsymbol{y}_{gijr} = (y_{gij1r}, y_{gij2r}, \ldots, y_{gijTr})'$ be the vector of gene expression values for gene $g$ across time points $\{1, \ldots, T\}$, and $\boldsymbol{\varepsilon}_{gijr} = (\varepsilon_{gij1r}, \varepsilon_{gij2r}, \ldots, \varepsilon_{gijTr})'$ be the vector of random errors across time ($i$ and $j$, as previously, denote the factor levels). Let $\pi_k = P(z_{gk} = 1)$ denote the a priori

membership probability for cluster $k$ ($\sum_{k=1}^{K} \pi_k = 1$). We then assume

$$P(\boldsymbol{y}_{gijr}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k P(\boldsymbol{y}_{gijr}|z_{gk}), \tag{1}$$

where, given $z_{gk} = 1$,

$$\boldsymbol{y}_{gijr} = \begin{pmatrix} \nu_{g1} + \alpha_{k1} i + \beta_{k1} j + \gamma_{k1} ij + \varepsilon_{gij1rk} \\ \nu_{g2} + \alpha_{k2} i + \beta_{k2} j + \gamma_{k2} ij + \varepsilon_{gij2rk} \\ \vdots \\ \nu_{gT} + \alpha_{kT} i + \beta_{kT} j + \gamma_{kT} ij + \varepsilon_{gijTrk} \end{pmatrix} = \boldsymbol{\nu}_g + \boldsymbol{\alpha}_k i + \boldsymbol{\beta}_k j + \boldsymbol{\gamma}_k ij + \boldsymbol{\varepsilon}_{gijrk},$$

$\boldsymbol{\nu}_g, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k$, and $\boldsymbol{\gamma}_k$ being the corresponding vectors of the model coefficients. From now on, we denote the error term as $\boldsymbol{\varepsilon}_{gijrk}$ to clarify its dependence on cluster membership of the gene $g$.

The experimental replicates are assumed to be independent, so that $\boldsymbol{\varepsilon}_{gijrk}$ ($r = 1, ..., R$) follows $N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, a multivariate normal distribution. The assumption of common variance $\sigma^2$ across genes and time points will be evaluated using residual analysis. The correlation matrix, $\boldsymbol{\Sigma}$, can be unstructured or can take on a structure reflecting the nature of correlation observed in the longitudinal data. For instance, if the correlation can be assumed to be constant over time, specified by a correlation coefficient $\rho$, one structure that can be assumed is the equicorrelation dispersion matrix, $\boldsymbol{\Sigma}(\rho) = [(1-\rho)\boldsymbol{I} + \rho\mathbf{1}\mathbf{1}']$, where $\boldsymbol{I}$ is an identity matrix of dimension $T$, and $\mathbf{1}$ is a $T$-dimensional vector of 1's. The model we propose here is a longitudinal mixture model in which $\boldsymbol{\nu}_g$ denotes the gene-specific baseline effect of gene $g$ (often representing the reference group), while the parameters $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ depend on the cluster membership $k$. Under this assumption, the genes are classified into clusters that share similar differential expression patterns, rather than their actual observed expression patterns, as the $\boldsymbol{\nu}_g$'s are different from gene to gene. Genes belonging to cluster $k$ will have identical factorial effects across time, but are allowed to vary individually and stochastically through the gene-specific baseline effect term. To generalize the model in (1) to multi-level factors, additional indicator variables will have to be defined for

the different levels of the factors and therefore additional model parameters including the main and interaction effects for each of the factorial combinations will need to be specified.

### *2.2. Likelihood and priors*

First, let us denote by $\boldsymbol{Z}$ the $G \times K$ matrix of the latent cluster indicator variables that take values $z_{gk}$ $(g = 1, \ldots, G;\ k = 1, \ldots, K)$. We also denote the set of all model parameters by $\boldsymbol{\eta}_1 = \{\boldsymbol{\pi}, \boldsymbol{\nu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \sigma^2\}$. We assume

$$P(\boldsymbol{y}_{gijr}|\boldsymbol{\eta}_1, z_{gk}) = \frac{1}{(2\pi)^{T/2}\sigma^T\,|\boldsymbol{\Sigma}(\rho)|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'_{gijrk}\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\varepsilon}_{gijrk}\right),$$

where $\boldsymbol{\varepsilon}_{gijrk}$ denotes $\boldsymbol{y}_{gijr} - \boldsymbol{\nu}_g - \boldsymbol{\alpha}_k i - \boldsymbol{\beta}_k j - \boldsymbol{\gamma}_k ij$. Then, with $\boldsymbol{Y}$ denoting the complete set of expression values $y_{gijtr}$ $(g = 1, \ldots, G;\ i, j \in \{0, 1\};\ t = 1, \ldots, T;\ r = 1, \ldots, R)$, the complete data likelihood function can be written as:

$$L(\boldsymbol{\eta}_1|\boldsymbol{Y}, \boldsymbol{Z}) = \prod_{k=1}^{K}\prod_{g=1}^{G}\prod_{i=0}^{1}\prod_{j=0}^{1}\prod_{r=1}^{R}[\pi_k P(\boldsymbol{y}_{gijr}|\boldsymbol{\eta}_1, z_{gk})]^{z_{gk}}. \tag{2}$$

Next, we elicit prior distributions for the model parameters, as below:

$$\begin{aligned}
\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)' &\sim \text{Dirichlet}(\theta_1, \ldots, \theta_K), \\
\boldsymbol{\nu}_g \sim \text{Normal}(\boldsymbol{m}, \sigma_\nu^2\kappa_g\boldsymbol{I}), &\quad \boldsymbol{m} \sim \text{Normal}(\boldsymbol{0}, 10^6\boldsymbol{I}), \\
\kappa_g \sim \text{Uniform}(l_g, u_g), (l_g > 0), &\quad \boldsymbol{\alpha}_k \sim \text{Normal}(\boldsymbol{0}, \sigma_\alpha^2\boldsymbol{I}), \\
\boldsymbol{\beta}_k \sim \text{Normal}(\boldsymbol{0}, \sigma_\beta^2\boldsymbol{I}), &\quad \boldsymbol{\gamma}_k \sim \text{Normal}(\boldsymbol{0}, \sigma_\gamma^2\boldsymbol{I}).
\end{aligned}$$

In the above, $g$ takes values in $\{1, \ldots, G\}$ and $k \in \{1, \ldots, K\}$ . Additionally, we assume, $\rho \sim \text{Uniform}(-1, 1)$, $P(\sigma^2) \propto \frac{1}{\sigma^2}$, $P(\sigma_\alpha^2) \propto \frac{1}{\sigma_\alpha^2}$, $P(\sigma_\beta^2) \propto \frac{1}{\sigma_\beta^2}$, $P(\sigma_\gamma^2) \propto \frac{1}{\sigma_\gamma^2}$, and $P(\sigma_\nu^2) \propto \frac{1}{\sigma_\nu^2}$. The additional set of prior hyperparameters is $\boldsymbol{\eta}_2 = (\boldsymbol{m}, \boldsymbol{\kappa}, \boldsymbol{l}, \boldsymbol{u}, \boldsymbol{\theta}, \sigma_\nu^2, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2)$, where $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_G)$, $\boldsymbol{l} = (l_1, \ldots, l_G)$, $\boldsymbol{u} = (u_1, \ldots, u_G)$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$.

*2.2.1. Choice of hyperpriors and prior hyperparameters.*

The Dirichlet $(\theta_1, \ldots, \theta_K)$ distribution is a conjugate prior density for $\boldsymbol{\pi}$, with a non-informative version having $\theta_k = 1$ for each $k$ $(k = 1, ..., K)$. $\boldsymbol{\nu}_g, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k$, and $\boldsymbol{\gamma}_k$ are assumed to follow independent multivariate normal distributions, while $\rho$ is assumed to have a Uniform prior. The parameter $\kappa_g$ allows genes to have different variances, irrespective of cluster membership. $l_g$ and $u_g$ can be chosen through an Empirical Bayes approach, or informed by biological knowledge. Repeated sampling of $\kappa_g$ in the model updating step can be computationally expensive, and was noticed in pilot runs to give no significant benefit. Hence in practice, we propose sampling $\kappa_g$ at the beginning of the run, and keeping it fixed through the MCMC procedure. The priors for variance parameters are chosen to be non-informative on the log-scale, as this appears more robust than a hierarchical inverse-gamma prior where inference can be highly sensitive to the choice of the inverse-gamma hyperparameters when the variance parameters are estimated to be close to zero [6]. The final set of parameters that should be updated in the course of the algorithm is now denoted as $\boldsymbol{\eta} = (\boldsymbol{\nu}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \sigma^2, \boldsymbol{m}, \sigma^2_\nu, \sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma)$. We next propose a Bayesian data augmentation approach in Section A.2 to estimate the model parameters and latent gene cluster memberships.

*2.2.2. A partially marginalized likelihood.*

In order to focus on the relevant model signals, that is, differential expression patterns determined by $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k$, and $\boldsymbol{\gamma}_k$, we consider all gene-specific baseline parameters $\boldsymbol{\nu}_g$'s as nuisance parameters and marginalize them out of the likelihood function as follows. Here, we let $\boldsymbol{\eta'} = \boldsymbol{\eta} \setminus \boldsymbol{\nu}$ which represents all parameters excluding $\boldsymbol{\nu} = \{\boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_G\}$. Then, with $\boldsymbol{\eta'} = \{\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \sigma^2, \boldsymbol{m}, \sigma^2_\nu, \sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma\}$, we have

$$L(\boldsymbol{\eta'}|\boldsymbol{Y}, \boldsymbol{Z}) = P(\boldsymbol{Y}|\boldsymbol{\eta'}, \boldsymbol{Z}) = \int \cdots \int P(\boldsymbol{Y}|\boldsymbol{\eta}, \boldsymbol{Z}) P(\boldsymbol{\nu}_1) \cdots P(\boldsymbol{\nu}_G) d\boldsymbol{\nu}_1 \cdots d\boldsymbol{\nu}_G, \quad (3)$$

After some algebraic manipulation, we have

$$P(\boldsymbol{Y}|\boldsymbol{\eta}', \boldsymbol{Z}) \propto \frac{|\boldsymbol{\Sigma}(\rho)|^{-2GR}}{\sigma^{(4R-1)GT}} \prod_{g=1}^{G} \left[ |\boldsymbol{\Lambda}_g|^{-1/2} \exp\left\{ -\frac{1}{2} \sum_{g=1}^{G} (A_g - B_g + C_g) \right\} \right], \text{ where} \tag{4}$$

$$A_g = \frac{1}{\sigma^2} \sum_{k=1}^{K} \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{r=1}^{R} \left( z_{gk} \boldsymbol{w}'_{gijrk} \boldsymbol{\Sigma}(\rho)^{-1} \boldsymbol{w}_{gijrk} \right),$$

$$B_g = \frac{1}{\sigma_g^2} \sum_{k=1}^{K} z_{gk} \boldsymbol{\mu}'_{gk} \boldsymbol{\Lambda}_g \boldsymbol{\mu}_{gk}, \qquad C_g = \frac{\boldsymbol{m}'\boldsymbol{m}}{\sigma_\nu^2 \kappa_g},$$

$$\boldsymbol{w}_{gijrk} = \boldsymbol{y}_{gijr} - \boldsymbol{\alpha}_k i - \boldsymbol{\beta}_k j - \boldsymbol{\gamma}_k ij,$$

$$\boldsymbol{\mu}_{gk} = \boldsymbol{\Lambda}_g^{-1} \left[ \frac{\sigma_g^2}{\sigma^2} \boldsymbol{\Sigma}(\rho)^{-1} \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{r=1}^{R} \boldsymbol{w}_{gijrk} + \sigma^2 \boldsymbol{m} \right],$$

$$\boldsymbol{\Lambda}_g = 4R\sigma_\nu^2 \kappa_g \boldsymbol{\Sigma}(\rho)^{-1} + \sigma^2 \boldsymbol{I}, \qquad \text{and } \sigma_g^2 = \sigma^2 \sigma_\nu^2 \kappa_g.$$

The marginalized model can be fitted using a hybrid Markov chain Monte Carlo algorithm. The steps involve sampling from the following distributions in turn: (i) $\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\eta}'$, (ii) sampling each of the parameters in $\boldsymbol{\eta}'$ either from their posterior full conditional distributions (Gibbs sampling) or marginal posterior distributions (Metropolis algorithm). Full details of the derivation of the marginalized likelihood, as well as the steps of the MCMC procedure, are given in Appendix A.

## 3. Simulation studies

We first assessed the impact of various model settings and assumptions on the performance of the proposed Bayesian longitudinal mixture model-based cluster discovery algorithm. These included: (*i*) different baseline variance or noise, $\sigma_\nu^2$, and the size of clusters, (*ii*) different (unknown) numbers of clusters, $K$, and (*iii*) exploration of assumptions of independence across time.

We generated several data sets with fixed numbers of clusters ($K = 2$ and 5), two binary factors, four time points ($T = 4$) and five replicates ($R = 5$). Table 1 shows the hypothetical model parameters, chosen to ensure a mix of gene expression patterns in the data, with and without interactions between factors. For simulating $K = 2$, the first two clusters were used, while for simulating $K = 5$, all five clusters were used. Figure 1 shows the temporal gene expression profiles of five representative genes, one from each cluster, using $\sigma_\nu = 0.5$. For cluster 1, plots 1 and 2 refer to the first level for factor 1 while plots 3 and 4 relate to the second level. The similarity of these two sets of plots indicate that gene expression does not differ between the two levels of factor 1, corresponding to $\boldsymbol{\alpha}_1$ having zero values for all time points. Expression profiles at level 1 for factor 2 (plots 1 and 3) are very different from expression profiles at level 2 of factor 2 (plots 2 and 4), which means $\boldsymbol{\beta}_1$ will have some non-zero values. Since the gene expression pattern across plots 1 and 2 at the first level of factor 1, is similar to that across plots 3 and 4 at the second level of factor 1, it indicates that there is no interaction between factors 1 and 2 in this cluster (i.e. $\boldsymbol{\gamma}_1 = 0$). For cluster 5, in contrast, gene expression profiles at two levels of one factor are not the same as in the two levels of the other factor and therefore, we will expect all model parameters ($\boldsymbol{\alpha}_5$, $\boldsymbol{\beta}_5$, $\boldsymbol{\gamma}_5$) to have some non-zero values. In this cluster, each factor interacts with and modifies the effect of the other in its gene expression pattern.

[*Table 1 about here*]

[*Figure 1 about here*]

**MCMC convergence assessment and posterior inference.** We used MCMC chains of at least 10,000 posterior simulations in fitting each model. Con-

vergence diagnostics were performed using the `R coda` package [21], to assess if the MCMC algorithms displayed sufficient mixing and convergence. These included graphical methods of trace-plots and autocorrelation plots for all the parameters to visually inspect the posterior distributions, as well as calculating numerical diagnostic criteria, including the Geweke [9] and Heidelberger-Welch tests [12], and the potential scale reduction factor (PSRF) [8]. The results of these diagnostics generally supported sufficient mixing and convergence in the posterior estimation of the model parameters. Although there is no definitive guideline on how much burn-in should be discarded, but in all cases a burn-in of 10% of the MCMC was supported by the range of the above diagnostic tests. A summary of MCMC convergence diagnostics is presented in Appendix C of the online Supplementary materials.

By specification, mixture components are a priori exchangeable across cluster labels, and although this does not necessarily lead to non-identifiability (at least in the Gaussian case) this may cause complications with the MCMC procedure in the form of label-switching. If the MCMC has mixed well enough, and with a large enough sample size, each component could end up with identical posterior distributions. In such a situation, the accurate computation of posterior summaries would depend on the ability to relabel the samples according to a specific component membership, which may be done by a post-processing step after observing the traceplots of the posterior samples [7]. This was carried out in all of our simulation studies and applications.

The performance of our method was evaluated using the averaged Mean Square Errors (MSE) for all the estimated parameters and the percentage of gene-to-cluster misclassification. The estimated posterior cluster membership of a gene was determined as the cluster for which the posterior probability for group membership was the highest (i.e., the mode). Next, we present the results of each simulation study.

### *3.1. Effect of noise parameters and cluster sizes*

Large variability in the baseline values for different genes could potentially decrease the precision of estimating the posterior densities due to larger noise-to-signal ratios. We therefore decided to conduct a study to determine whether the quality of posterior estimation was worsened, when the variation in gene baseline values was increased.

For every gene $g$ at any time point $t$, the gene-specific baseline effect terms $\boldsymbol{\nu}_g$'s were simulated from $N(\mathbf{0}, \sigma_\nu^2 \boldsymbol{I})$, with $\sigma_\nu$ taking on three different values: $0.5, 2$ and $5$, while the error terms, $\boldsymbol{\varepsilon}_{gijr}$'s, were simulated from $N(\mathbf{0}, \boldsymbol{I})$ where $\sigma^2 = 1$. We simulated data sets for two gene set sizes ($G = 100, 300$), as well as balanced (same number of genes per cluster) and unbalanced (different number of genes per cluster) clusters. The hyperparameters $\sigma_\alpha^2, \sigma_\beta^2$, and $\sigma_\gamma^2$, were set to be $10^3$, to be weakly informative. The results for this part are summarized in Table 2. It can be seen that, under different simulation scenarios, the algorithm was able to correctly identify all the gene clusters and classify the genes. In the posterior estimation of the model coefficients, namely, $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, the MSEs in each case remain low, and there is no increasing pattern with an increase in $\sigma_\nu^2$. Increasing $\sigma_\nu^2$ did not, therefore, seem to have a negative effect on the posterior estimation of the model coefficients, at least within the range of variation of $\sigma_\nu^2$ considered.

[*Table 2 about here*]

### *3.2. Model assessment*

In a real experiment we typically do not have any prior knowledge of the number of clusters. In principle, one may set up a reversible jump MCMC algorithm [10] to jointly sample $K$ and $\boldsymbol{\eta}$, but the complexity of the model would make it difficult to design efficient proposal densities, thus adding to an already high computational cost. Instead, we take a criterion-based approach to assess the performance of our method under different assumed numbers of clusters. In a Bayesian framework, the Bayes Factor is the natural criterion for model selection (in this case, the number of clusters); but this would require (i) the marginal likelihood to be computed in closed form, and (ii) summing the marginal likelihoods over all possible gene-cluster allocations, which would be computationally prohibitive. Instead, we used an alternative Bayesian criterion, the Deviance Information Criterion (DIC), defined as

$$DIC = 2E\left[D(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\eta}')|\boldsymbol{Y}, \boldsymbol{Z}\right] - D(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\eta}'),$$

where $D(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\eta}') = -2 \log L(\boldsymbol{\eta}' | \boldsymbol{Y}, \boldsymbol{Z})$ is the deviance [29]. The first term is the posterior expected deviance and is estimated by

$$\frac{1}{S} \sum_{s=1}^{S} D(\boldsymbol{Y}, \boldsymbol{Z}^{(s)}, \boldsymbol{\eta}'^{(s)}),$$

calculated using posterior sampling, while the second term is estimated by $D(\boldsymbol{Y}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{\eta}}')$ where $\hat{\boldsymbol{Z}}$ is the $G \times K$ matrix constructed using the posterior modes of gene assignment and $\hat{\boldsymbol{\eta}}'$ represents the posterior means of the model parameters excluding $\boldsymbol{\nu}$. The selected model should correspond to the lowest value of the DIC.

To assess the ability of our method in finding the appropriate number of clusters, we carried out a simulation study varying the true total number of clusters in the data, and allowing model fitting over a range of cluster counts. We simulated two data sets: (i) a smaller with $K = 2$ ($G = 100$, balanced) and (ii) a larger $K = 5$ ($G = 200$, unbalanced) clusters, with $\sigma_\nu = 0.5$ in each case. For each data set, we applied our algorithm assuming total cluster counts of $2, 3, 4, 5$, and 6. The goal was to assess if the DIC consistently indicated the selection of the correct number of clusters, in spite of being allowed to fit a range of models with varying numbers of clusters.

Results for this study are presented in Table 3. In the table, $K$ refers to the actual number of gene clusters in the data, "Assumed $K$" refers to the maximum number of clusters that the method was allowed to fit in a specific run, and $\hat{K}$ refers to the actual number of non-empty clusters found after the MCMC algorithm to fit the model was run. (An empty cluster was defined as one for which no data point had a posterior probability of membership greater than 0.01.) For the simulated data set with an actual value of $K = 2$, it was interesting to see that even when a larger number of clusters was allowed (upto an assumed $K = 6$), there were never more than $\hat{K} = 2$ non-empty clusters found, indicating the robustness of the model. Sampling fluctuations in the parameters from their posterior distribution caused some slight variation in the value of the DIC, between 31773 to 31827, but the MSE only varied between 0.017 to 0.018, with 0% misclassification. For the simulated data with true $K = 5$, when we set an assumed $K < 5$, the algorithm merged some of the clusters together and led to a high rate of misclassification and a high value of the DIC. The best performance was when

allowing for an assumed $K = 5$ and 6, in both of which cases the algorithm identified a correct total of 5 clusters; again showing that a larger model than necessary was not chosen (the sixth cluster being empty). The DIC was minimized for $\hat{K} = 5$, with the MSE varying between 0.032 and 0.033 and a 0% misclassification rate. Results were consistent when multiple runs were done with any set value of $K$. The results from Table 3 indicate that in a real-life scenario, since the algorithm tends not to overestimate the true number of clusters, one may consider running multiple MCMC chains of the algorithm for different assumed numbers of clusters, and choose $\hat{K}$ based on the lowest possible value of the DIC that gives consistent results over multiple runs.

[*Table 3 about here*]

### *3.3. Exploring dependence structure and a comparative study*

We next examined if the cluster discovery algorithm continued to perform well if we ignored the correlation structure. We used the algorithm, under an assumption of independent time points, on a data set simulated with gene expression values which were correlated over time. In this study, we considered two settings: balanced and unbalanced genes per cluster and under each setting, we simulated three gene expression data sets assuming an auto-regressive correlation structure with correlation parameter taking on three different values, $\rho = 0.1, 0.5$, and 0.9. For these six data sets, we assumed $K = 5$ and $\sigma_\nu = 0.5$.

Additionally, we wanted to compare the proposed Bayesian longitudinal mixture model and cluster discovery algorithm with the widely used model-based clustering method MCLUST [4]. MCLUST does not model the factorial time-course structure of the gene expression measurements, so an adaptation of the data was made to make the results more comparable. Before MCLUST was run, the six data sets in the previous part were re-organized as follows. We calculated the mean expression values using the values of five replicates, and then the differences between the means of other groups (i.e. $(i, j) = \{(0, 1), (1, 0), (1, 1)\}$) and the mean of the reference group (i.e. $(i, j) = (0, 0)$) across the time points. As a result, the mean differences in observed expression values of one gene were considered as one single vector, $\overline{\boldsymbol{y}}_{gij.} - \overline{\boldsymbol{y}}_{g00.}$. MCLUST assumes that

the vector of expression values, $\overline{\boldsymbol{y}}_{gij\cdot} - \overline{\boldsymbol{y}}_{g00\cdot}$, follows a mixture of multivariate Gaussian distributions. It then chooses among different structures for the variance-covariance matrix of the $k$th cluster, $\boldsymbol{\Sigma}_k$ representing the geometric structure of the cluster surface such as ellipsoidal or spherical ($\boldsymbol{\Sigma}_k = \lambda \boldsymbol{I}$), ranging from equal variance across clusters $\boldsymbol{\Sigma}$ to unconstrained variances. The EM algorithm is employed to estimate parameters and cluster memberships, and the Bayesian Information Criterion (BIC) is used for model selection or determining the number of clusters along with different variance-covariance structures. We specified the number of clusters as between 1 and 9 and the best model according to the BIC profiles of different variance-covariance structures was chosen.

The results, comparing our proposed method and MCLUST, are summarized in Table 4, suggesting that even when our algorithm ignores correlation between the time points, it can still correctly identify the true number of clusters in each case. We also can see that MCLUST performed comparably for this simulation setting. However, with an increase in the volume and dimensionality of the data, the EM-based clustering algorithm in MCLUST is likely to be challenged, as noted by Fraley and Raftery [4]. Also, by compressing the data before application of MCLUST, we lose valuable information and the ability to estimate the degree of interaction between factors in the different clusters, so it is not ideal for this setting. Next, we applied our methods to the original motivating data set on bone aging in mice, discussed in Section 4.

[*Table 4 about here*]

## 4. Application to Bone Aging Study in Mice

A large gene expression data set was compiled in a study at the Orthopaedic Research Laboratory at Boston University School of Medicine, that assessed bone aging in male and female mice expressing one or both copies of the Brd2 transcription factor. In these studies bone mass was shown to be lower in females than males, and loss was associated with aging. Bone loss in the Brd2 heterozygote female mice was greater than seen in the wild type females, but in males, the temporal pattern of bone loss

was similar between the two genotypes. The experiment was conducted using a $2^2$ factorial design (sex: male or female; genotype: wild-type or heterozygous/mutant); mice were randomly assigned into the four groups. Data on the mice in each group were analyzed at 3, 6, 9, and 12 months. At each time point, humeri were harvested from four mice per group, sacrificing the mice. The bones were used to extract total RNAs, which were used for microarray analyses. Concurrently, tibia from the same animals were analyzed to assess the changes in bone quality as a function of sex and phenotype. Thus, different mice for each of the four groups were followed for each of the four time points, giving a total of 64 mice. Since the potential for serial correlation was small here, we assumed an identity covariance structure for the error model, $\boldsymbol{\Sigma}(\rho) = \boldsymbol{I}$. The next goal was to test the performance of our clustering method and to differentiate the effects of sex and reduced expression of Brd2 on the global gene expression patterns as the animals aged. The data pre-processing steps are detailed in Appendix B in the online Supplementary Materials.

We applied the gene cluster discovery algorithm programmed in `R` [23] on the final set of 3,950 genes under the LinGA computing cluster on the Boston University Medical Campus. Pilot runs of the algorithm, varying the total number of gene clusters between 20 and 50, suggested that the number of clusters was likely to be between 18 and 25. The hyperparameter choices were taken to be the same as in Section 3.1, after pilot runs varying these settings across a wide range indicated no major differences in the results. For our final analysis, two independent runs of the algorithm were conducted with two different initial numbers of gene clusters: 25 and 30, for 10,000 MCMC iterations in which 21 and 22 non-empty clusters were reported at the end, respectively, with overall DIC values $\approx 3.11 \times 10^8$ and $3.12 \times 10^8$ respectively. Results were based on the last 90% of the MCMC chains. Standard MCMC convergence diagnostics did not indicate any issues with lack of convergence of the simulations. (Details of convergence diagnostics are provided in Appendix C of the online Supplementary materials.) After comparing the results of the two applications, we identified 18 gene clusters that were consistently estimated between these two runs, displayed in Figure 2. The sex effect is represented by $\boldsymbol{\alpha}_k$ while the Brd2 gene effect by $\boldsymbol{\beta}_k$, and $\boldsymbol{\gamma}_k$ is the interaction of these two factors. The posterior distributions of the model parameters

of these 18 clusters can be found in Supplementary Tables D1 and D2. Both post-hoc residual plots and normal quantile-quantile plots suggested the assumptions of normality and common variance across genes and time points were sufficient. We also carried out several posterior predictive checks on the model fit along the lines suggested in Gelman et al. [7], comparing a number of features in the data which were not directly estimated in the model (such as extreme values, harmonic means, skewness, kurtosis and mean absolute deviation, to name a few) between datasets generated from the the posterior predictive distribution and the original data. These gave no strong indication of the lack of model fit to the data, with posterior predictive P-values in the range of 0.45 to 0.74.

Notable differences were observed through qualitative comparison of the characteristics of each cluster (Figure 2). The graphical patterns of the experimental groups in cluster 14 indicated that differences between male and female mice were more pronounced compared to genotype-based differences within each sex. In contrast, cluster 4 indicated that the set of genes in this cluster had similar expression levels within the male mice. The expression patterns for the female heterozygous mice were different from the wild-type female mice but resembled that of the male mice. Cluster 8 showed that there are no differences between female wild-type and heterozygous mice until 9 months of age after which the gene expression levels diverge. This qualitative assessment allowed us to identify gene groups that showed commonality by sex or genetic phenotype as well as how genotype might interact with sex.

### *4.1. Comparison with mixture model-based inference*

As in the simulation studies, we applied the MCLUST algorithm on the mouse data set to compare the inference between methods. A similar adaptation of the data set was made as discussed in Section 3.3. MCLUST was allowed to run with the number of clusters set to be between 5 and 30, and a choice of all 14 models with different covariance structures that were available. Using BIC as the model selection criterion, a model with 11 clusters, and ellipsoidal, equal shape and orientation (VEE) was chosen. Supplementary Figure D2 (online) shows the gene expression profiles by cluster, tabulated across time points and factor combinations. Unlike the clusters found by

the factorial time-course model as shown in Figure 2, these clusters do not show a very clear difference in profiles (or effects of the interaction between time and factor levels) between clusters. In fact, for a number of clusters (e.g. Clusters 1, 2, 4, 5, 6) the profiles for certain factor combinations look almost identical, with some variations in the magnitude. The adjusted Rand index [13] between MCLUST and the Bayesian longitudinal mixture model-based clusterings is 0.1222, indicating a minor overlap, as also demonstrated in the biplot comparing the two clusterings (Supplementary Figure D3).

[*Figure 2 about here*]

### *4.2. Scientific exploration of model-fitting results*

The next point in our investigation was to determine how far the results from the factorial time course gene cluster discovery algorithm could take us into gaining biological insights into the underlying genetics of bone aging. In order to assess the biological relevance of the 18 gene clusters found by our method, the gene expression data was analyzed using QIAGEN's Ingenuity® Pathway Analysis tool (`http://www.qiagen.com/ingenuity`, IPA®, QIAGEN Redwood City), henceforth termed IPA, in an effort to identify all biological functions of genes in each cluster. These biological functions were then assigned to more than 80 categories of general functions. Further, grouping of these categories was performed in order to present the data in terms of broader functional groupings such as skeletal, immune, and inflammatory related functions that would be relatable to the known phenotypic differences of the four groups of animals (Supplementary Table D3). The miscellaneous category contains functions such as hereditary disorder, ophthalmic disease and other diseases that are not directly related to bone aging. Figure 3 presents a bar plot of all biological functions for each cluster as well as the set of all genes used in the gene cluster discovery algorithm. As visualized in the bar plot, about 30% of all genes related to hematological/immunological functions were differentially expressed across the clusters (i.e. dark blue to light blue), but on the other end of the spectrum, there were about 10% of all genes related to skeletal/muscular tissue functions (i.e. dark brown).

This was consistent to previous findings on known and predicted phenotypes [2].

[*Figure 3 about here*]

Since the primary focus of this study was to assess the sex-linked effect of Brd2 on bone aging, skeletal-related biological functions were analyzed further. Immune and inflammatory related biological functions were alternatively examined in order to compare the results to existing literature on the effects of Brd2 mutations on the immune system. Clusters 4, 14, and 18 had the highest percentages of genes that were identified by IPA as musculoskeletal, with cluster 4 showing the largest number of genes that had previously been identified in bone tissues. Clusters 1, 2, 8 and 12 had the highest percentages of genes associated with immune and inflammatory functions, and there is evidence that Brd2 is involved in regulation of such genes [1]. We chose clusters 4 and 12 for further examination. The pattern observed in cluster 4 would suggest that the effect of the mutation of the Brd2 gene would make the female heterozygous animals show an age-related pattern of gene expression similar to that seen in both strains of male animals. The top ranked diseases and biological functions, and canonical pathways in Clusters 4 and 12 are listed in Supplementary Table D4.

Cluster 4 showed a significant interaction between the genotype and sex. This indicates that over time, gene expression patterns in the Brd2 mutant female mice resemble those in male mice (Figure 2). This was also observed when looking at network interactions predicted by the IPA software (Figure 4). For example, for metabolic bone disease, all genes in this cluster were down-regulated in male (Brd2 mutant and wild-type) and female Brd2 mutant mice but not in female wild-type mice. These results are also in agreement with quantitative data on bone loss in the trabecular bone compartments of the tibia and the vertebrae of these same groups since the affected genes in this group are all known to express proteins that are physical components of the extracellular matrix of bone. Brd2 has no effect on bone loss in male mice over time, whereas in female mice, the altered Brd2 expression resulted in more bone loss over time compared to wild-type in a manner similar to male mice although the interaction was only seen at later time points and was less in the vertebral bone (Supplementary Figure D4). The results indicated that even though females have a

decreased trabecular bone volume fraction compared to males, their bone metabolism was increased.

[*Figure 4 about here*]

Cluster 12 contained a high number of inflammation-related genes. When comparing to female wild-type mice, all male mice, both wild-type and mutant, showed some differences in gene expression patterns. On the other hand, female mutant mice demonstrated more divergent differences in gene expression. These sex-related differences are in agreement with the observed biological differences between male and female response to infection of the mammalia (Figure 4). Together these results suggested that the increased metabolic activity in females renders them more susceptible to adipogenesis (formation of fatty tissue).

## 5. Discussion

Dynamic microarray experiments are particularly suitable to help biologists understand the interwoven processes in which biological functions take place over a time period. We have proposed a Bayesian statistical framework for longitudinal mixtures of distributions and developed an algorithm to simultaneously estimate the model parameters and discover gene clusters that reflect biological networks and pathways within functioning cells. The estimated parameters and gene clusters can be easily post-processed using appropriate software to further reveal clusters enriched in certain biological functional networks. We have demonstrated the performance of this proposed model and algorithm in both synthetic and real data sets, showing accurate parameter estimation and a biologically meaningful classification of genes.

When modeling data from the factorial design, we have chosen to adapt a linear modeling approach to capture the differential expression signals between levels of factors while allowing each gene to have its own random baseline effect. Comparing to other modeling approaches such as hidden Markov models or non-parametric cubic spline methods, the simplicity of the linear model allows for parsimony and offers easy interpretation of the estimated parameters. The model also specifies the interaction

effect between two factors and the significance of these interaction effects can be assessed longitudinally. Conceptually, the model framework we propose in this paper can be theoretically extended to include more than two factors and factors having more than two levels. As the model becomes more complex, the number of parameters of interest will also increase, making posterior simulation more challenging and computationally intensive. However, with the advent of highly powerful computing servers, the speeds of computation can be drastically improved. The Bayesian choice makes it easier for investigators to incorporate prior information on the temporal expression profiles if previous studies demonstrated that certain profile shapes are plausible.

One difficulty in mixture modeling approaches for gene expression is inferring the number of existing clusters in the data set, implicitly assuming that these clusters and their assigned genes are stable during the time-frame of the study. It is possible that some genes are considered *clusterless* as they may share features of multiple clusters or they do not belong to any cluster. We have shown earlier that the proposed modeling approach and cluster discovery algorithm do not tend to over-estimate the number of clusters in both simulated and real data. The deviance information criterion (DIC) could be used to choose the appropriate number of clusters.

An important objective of modeling gene expression data is to reveal latent functional pathways through cluster identification and analysis. Our proposed model was shown to have the ability to determine groups of genes which had differentiated patterns of expression between males and females over time, which led to a number of interesting conclusions regarding the bone biology. In Brd2 mutant mice, more bone loss was observed in females over time as compared to males, and a cluster of genes that were potentially responsible for this difference could be determined. With the increasing availability of diverse types of genomic and epigenetic data from high-throughput experiments as well as the huge growth in annotational databases, our proposed framework can be adapted to combine different sources of information to yield deeper insights into biological pathways. It can also be adapted to infer expression clusters in longitudinally measured factorial gene expression experiments measured through high-throughput sequencing, technology for which is becoming more widely available.

## 6. Software

Software in the form of R code, along with data sets and documentation, are available on request from the corresponding author.


## Acknowledgements

This research was supported by the National Institute of General Medical Sciences of the National Institute of Health under award number T32GM074905. It was conducted using the Linux Clusters for Genetic Analysis (LinGA) computing resources at the Boston University Medical Campus. The work on bone aging was supported by an Ellison Foundation grant and NIH NIAMS 1RO1AR05974 to LCG, and CTSA award UL1-TR000157 which funds the Microarray facility at BUSM. The authors would like to thank Drs. Elise Morgan, Gerard Denis, Darlene Lu and Serkalem Demissie who were responsible for the data generation through bone harvesting and microCT analysis, as well as Ms. Faye Andrews, Ms. Danielle Salazar and Ms. Dana Daukss for their technical assistance. The authors are also grateful to two anonymous referees whose comments led to a much improved version of the final manuscript.


## References


[1] G.P. Andrieu, J. S. Shafran, J. T. Deeney, K. R. Bharadwaj, A. Rangarajan and G. V. Denis, *BET proteins in abnormal metabolism, inflammation, and the breast cancer microenvironment.* J Leukoc Biol. 104 (2018), pp. 265-274.

[2] A. C. Belkina, , B. S. Nikolajczyk and G. V. Denis, *BET protein function is required for inflammation: Brd2 genetic disruption and BET inhibitor JQ1 impair mouse macrophage inflammatory responses.* J. Immunol. 190 (2013), pp. 3670–3678.

[3] J. Ernst, G. J. Nau and Z. Bar-Joseph, *Clustering short time series gene expression data.* Bioinformatics 21 (2005), pp. i159–i168.

[4] C. Fraley and A. E. Raftery, *Model-based clustering, discriminant analysis, and density estimation.* J. Amer. Stat. Assoc. 97 (2002), pp. 611–631.

[5] P. Gabbur, J. Hoying and K. Barnard, *Multimodal probabilistic generative models for time-*

*course gene expression data and Gene Ontology (GO) tags*. Math. Biosci. 268 (2015), pp. 80–91.

[6] A. Gelman, *Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)*. Bayesian Anal. 1 (2006), pp. 515–534.

[7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis*. Texts in Statistical Science Series, Chapman & Hall, London, 3rd edition.

[8] A. Gelman and D. B. Rubin, *Inference from iterative simulation using multiple sequences*. Statist. Sci., 7 (1992), pp. 457–472.

[9] J. F. Geweke, *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Bayesian Statistics 4 (1992), pp. 169–193.

[10] P. J. Green, *Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination*. Biometrika, 82 (1995), pp. 711–732.

[11] M. Gupta, P. Qu and J. G. Ibrahim, *A temporal hidden Markov regression model for the analysis of gene regulatory networks*. Biostatistics 8 (2007), pp. 805–820.

[12] P. Heidelberger and P. Welch, *Simulation run length control in the presence of an initial transient*. Oper. Res. 31(6) (1983), pp. 1109–1144.

[13] L. Hubert,and P. Arabie, *Comparing partitions*. J. Classif. 2 (1985), pp. 193-218.

[14] A. I. Hussein, J. Wu, M. Gupta and L. Gerstenfeld. *A Bayesian Approach to Assess the Transcriptome of Bone Aging and the Role of the Brd2 Gene in the Regulation of Sex Linked Bone Loss*. Presented at the ORS 2015 Annual Meeting, Las Vegas, Nevada.

[15] M. Kayano, H. Matsui, R. Yamaguchi, S. Imoto and S. Miyano, *Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection*. Biostatistics 17 (2016), pp. 235–248.

[16] E. de Klerk, J. T. den Dunnen, and P. A. t Hoen. *RNA sequencing: from tag-based profiling to resolving complete transcript structure*. Cell. Mol. Life Sci. 71 (2014), pp. 3537–3551.

[17] T. L. Lenstra, J. Rodriguez, H. Chen, and D. R. Larson. *Transcription Dynamics in Living Cells*. Annu Rev Biophys. 45 (2016), pp. 25–47.

[18] Y. Li, Y. He and Y. Zhang, *Analyzing gene expression time-course based on multi-resolution shape mixture model*. Math. Biosci. 281 (2016), pp. 74–81.

[19] X. Lu and Y. Huang, *Bayesian analysis of nonlinear mixed-effects mixture models for*

*longitudinal data with heterogeneity and skewness*. Statist. Med 33 (2014), pp. 2830–2849.

[20] P. Ma, C. I. Castillo-Davis, W. Zhong, and J. S. Liu, *A data-driven clustering method for time course gene expression data*. Nucleic Acids Res.34 (2006), pp. 1261–1269.

[21] M. Plummer, N. Best, K. Cowles and K. Vines, *CODA: convergence diagnosis and output analysis for MCMC*. R News 6(1) (2006), pp. 7-11.

[22] J. Quackenbush. *Computational approaches to analysis of DNA microarray data*. Yearb Med Inform. (2006), pp. 91–103.

[23] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2016, Vienna, Austria. ISBN 3-900051-07-0.

[24] A. Raj and A. van Oudenaarden. *Nature, nurture, or chance: stochastic gene expression and its consequences*. Cell 135 (2008), pp. 216–226.

[25] M. F. Ramoni, P. Sebastiani and I. S. Kohane, *Cluster analysis of gene expression dynamics*. Proc. Natl. Acad. Sci. U.S.A. 99 (2002), pp. 9121–9126.

[26] C. Sabatti and G. M. James, *Bayesian sparse hidden components analysis for transcription regulation networks*. Bioinformatics 22 (2006), pp. 739–746.

[27] A. Schliep, A. Schonhuth and C. Steinhoff, *Using hidden Markov models to analyze gene expression time course data*. Bioinformatics 19 (2003) , pp. i255–263.

[28] D. Scholtens, A. Miron, F. M. Merchant, A. Miller, P. L. Miron, J. D. Iglehart and R. Gentleman, *Analyzing factorial designed microarray experiments*. J. Mult. Anal. 90 (2004), pp. 19–43.

[29] D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. Van Der Linde, *Bayesian measures of model complexity and fit*. J. Royal Statist. Soc. B 64 (2002), pp. 583–639.

[30] D. A. Stavreva, L. Varticovski, and G. L. Hager. *Complex dynamics of transcription regulation*. Biochim. Biophys. Acta 1819 (2012), pp. 657–666.

[31] K. Viele and B. Tong, *Modeling with mixtures of linear regressions*. Statist. Comput. 12 (2002), pp. 315-330.

[32] J. Wang. *Computational biology of genome expression and regulation–a review of microarray bioinformatics*. J. Environ. Pathol. Toxicol. Oncol. 27 (2008), pp. 157–179.

[33] K. Wang, S. K. Ng and G. J. McLachlan, *Clustering of time-course gene expression profiles using normal mixture models with autoregressive random effects*. BMC Bioinformatics 13 (2012), pp. 300.

[34] F. Wang, H. Liu, W. P. Blanton, A. Belkina, N. K. Lebrasseur and G. V. Denis, *Brd2 disruption in mice causes severe obesity without Type 2 diabetes*. Biochem. J. 425 (2010),

pp. 71–83.

[35] Z. Wei and H. Li, *A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data*. Ann. Appl. Stat. 2 (2008) pp. 408–429.

[36] M. Yuan and C. Kendziorski, *Hidden Markov models for microarray time course data in multiple biological conditions*. J. Amer. Statist. Assoc. 101 (2006), pp. 1323–1332.

[37] Y. Zeng and J. Garcia-Frias, *A novel HMM-based clustering algorithm for the analysis of gene expression time-course data*. Comput. Stat. Data Anal. 50 (2006), pp. 2472–2494.

[38] B. Zhou and W. H. Wong, *A bootstrap-based non-parametric ANOVA method with applications to factorial microarray data*. Statistica Sinica 21 (2011), pp. 495–514.

[39] B. Zhou, W. Xu, D. Herndon, R. Tompkins, R. Davis, W. Xiao, W. H. Wong, and Inflammation and Host Response to Injury Program, *Analysis of factorial time-course microarrays with application to a clinical study of burn injury*. Proc. Natl. Acad. Sci. U.S.A. 107 (2010), pp. 9923–9928.

[40] J. Zierer, C. Menni, G. Kastenmuller, and T. D. Spector. *Integration of 'omics' data in aging research: from biomarkers to systems biology*. Aging Cell 14 (2015), pp. 933–944.

**Table 1.** Model parameters selected for simulation study.

| $k$ | $\boldsymbol{\nu}_g$ | $\alpha_{k1}$ | $\beta_{k1}$ | $\gamma_{k1}$ | $\alpha_{k2}$ | $\beta_{k2}$ | $\gamma_{k2}$ | $\alpha_{k3}$ | $\beta_{k3}$ | $\gamma_{k3}$ | $\alpha_{k4}$ | $\beta_{k4}$ | $\gamma_{k4}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $N(\mathbf{0}, \sigma_\nu^2 \boldsymbol{I})$ | 0 | -9 | 0 | 0 | -2 | 0 | 0 | 5 | 0 | 0 | 1 | 0 |
| 2 | $N(\mathbf{0}, \sigma_\nu^2 \boldsymbol{I})$ | -9 | -9 | 19 | -6 | -6 | 13 | 0 | -2 | 0 | 0 | 0 | 0 |
| 3 | $N(\mathbf{0}, \sigma_\nu^2 \boldsymbol{I})$ | 3 | 3 | -1 | 3 | -6 | -3 | 4 | 7 | -4 | -1 | 5 | 4 |
| 4 | $N(\mathbf{0}, \sigma_\nu^2 \boldsymbol{I})$ | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 3 | -5 | 0 |
| 5 | $N(\mathbf{0}, \sigma_\nu^2 \boldsymbol{I})$ | 5 | 7 | 3 | 4 | -1 | 5 | 4 | 3 | 2 | -5 | 0 | 0 |

**Table 2.** Performance of the proposed Bayesian longitudinal mixture model and cluster discovery algorithm for different values of $\sigma_\nu$ and different numbers of clusters, $K$, for both balanced and unbalanced clusters. Values given correspond to the $MSE$, and the percentage of misclassified genes.

| | $K = 2$ | | $K = 5$ | |
|---|---|---|---|---|
| | $G = 100$ Balanced | $G = 300$ Unbalanced | $G = 100$ Balanced | $G = 300$ Unbalanced |
| $\sigma_\nu$ | MSE, % | MSE, % | MSE, % | MSE, % |
| 0.5 | 0.0294, 0% | 0.0193, 0% | 0.0502, 0% | 0.0187, 0% |
| 2.0 | 0.0380, 0% | 0.0181, 0% | 0.0513, 0% | 0.0207, 0% |
| 5.0 | 0.0289, 0% | 0.0167, 0% | 0.0601, 0% | 0.0242, 0% |

**Table 3.** Performance of the proposed Bayesian longitudinal mixture model and cluster discovery algorithm for different assumed numbers of clusters. Numbers given correspond to the DIC, estimated number of clusters $\widehat{K}$, MSE, and percentage of misclassified genes in each case.

| | $K = 2$ ($G = 100$, Balanced) | $K = 5$ ($G = 200$, Unbalanced) |
|---|---|---|
| Assumed $K$ | DIC, $\widehat{K}$, MSE, % | DIC, $\widehat{K}$, MSE, % |
| 2 | 31827, 2, 0.018, 0% | 98473, 2, 6.085, 55% |
| 3 | 31777, 2, 0.017, 0% | 98391, 3, 5.991, 35% |
| 4 | 31809, 2, 0.017, 0% | 82675, 4, 3.840, 18% |
| 5 | 31878, 2, 0.017, 0% | 63041, 5, 0.032, 0% |
| 6 | 31773, 2, 0.018, 0% | 63143, 5, 0.033, 0% |

**Table 4.** Performance of the proposed Bayesian longitudinal mixture model and cluster discovery algorithm ignoring dependency between the time points for longitudinal gene expression data simulated under autoregressive correlation structure with $\rho = 0.1, 0.5, 0.9$ and $K = 5$. Comparison is made with the MCLUST algorithm [4]. No genes were misclassified in any of the simulations here.

| | | $\rho = 0.1$ | $\rho = 0.5$ | $\rho = 0.9$ |
|---|---|---|---|---|
| | Methods | $\widehat{K}$, MSE | $\widehat{K}$, MSE | $\widehat{K}$, MSE |
| $G = 100$, Balanced | Proposed | 5, 0.057 | 5, 0.046 | 5, 0.040 |
| | MCLUST | 5, 0.022 | 5, 0.018 | 5, 0.013 |
| $G = 200$, Unbalanced | Proposed | 5, 0.028 | 5, 0.033 | 5, 0.021 |
| | MCLUST | 5, 0.012 | 5, 0.011 | 5, 0.019 |

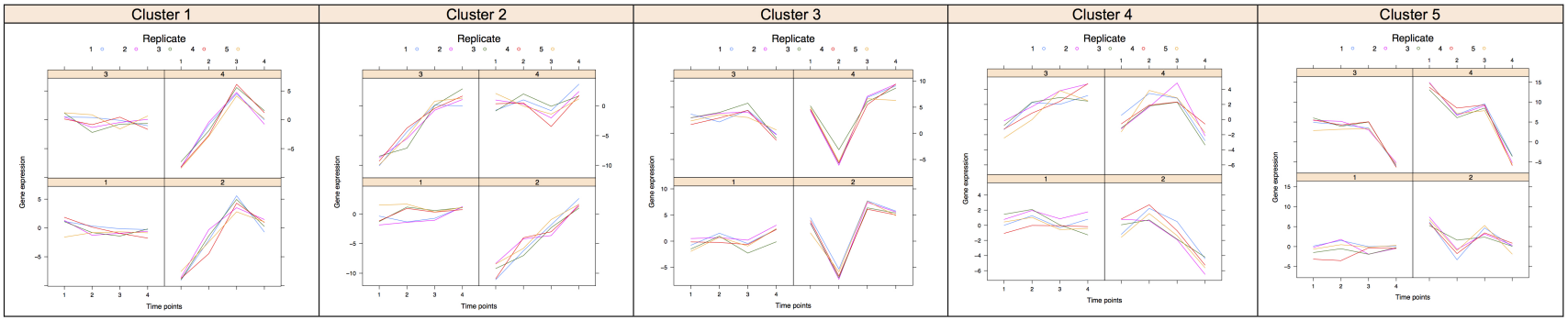


**Figure 1.** Factorial time-course gene expression profiles for five simulated clusters from Table 1, $k$=1 to 5. For each cluster, panel 1 (bottom left) refers to female wild-type as reference group; panel 2 (bottom right) to female heterozygote (i.e. mutant); panel 3 (top left) to male wild-type; and panel 4 (top right) to male heterozygote. Vertical axes of the plots are not drawn to the same scale.

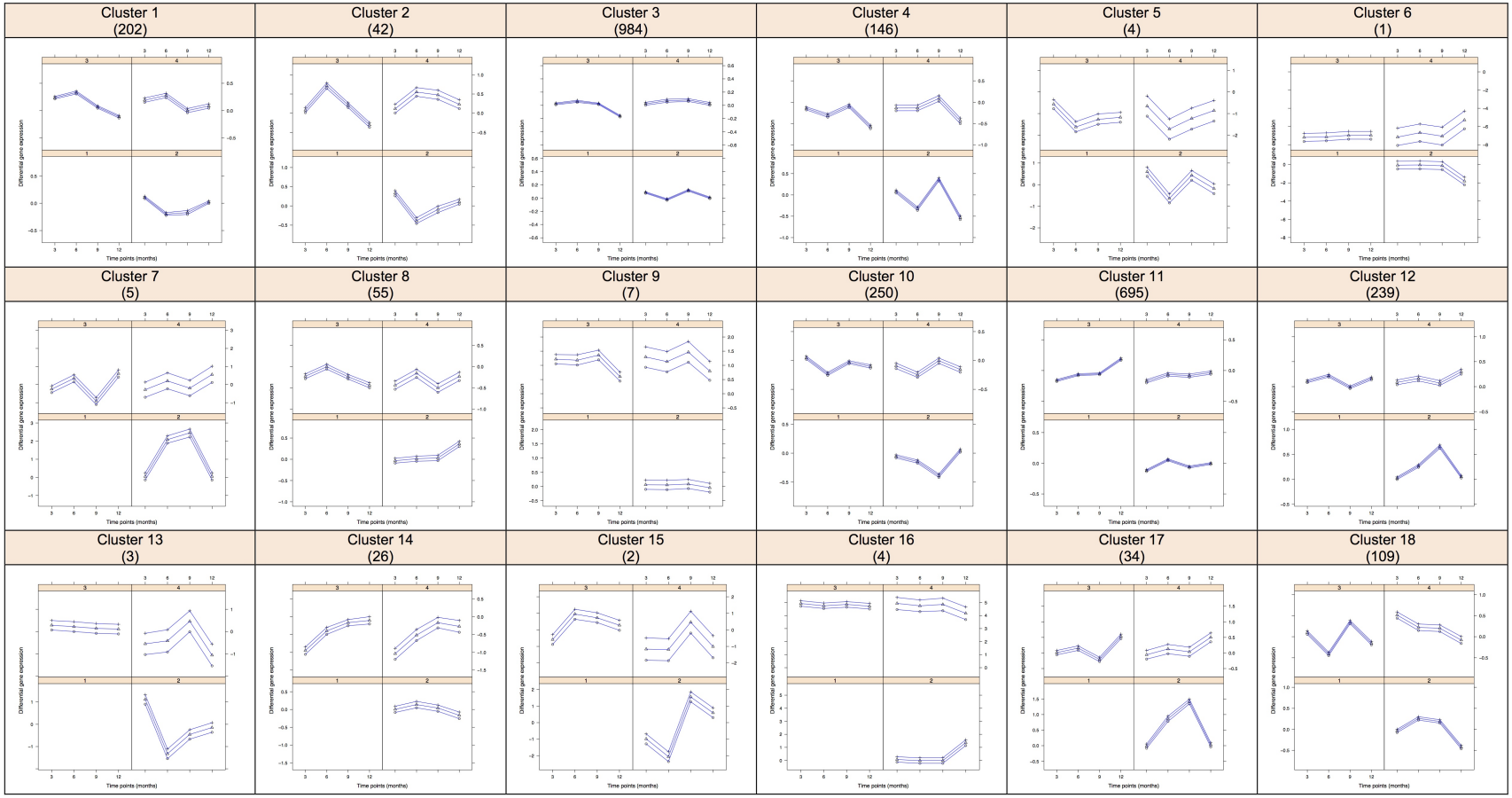


**Figure 2.** Eighteen clusters of factorial time-course differential gene expression patterns (compared to reference group) consistently estimated in two applications. For each cluster, panel 1 (bottom left) refers to female wild-type as reference group (not plotted); panel 2 (bottom right) refers to female heterozygote ($\boldsymbol{\beta}_k$); panel 3 (top left) refers to male wild-type ($\boldsymbol{\alpha}_k$); and panel 4 (top right) refers to male heterozygote ($\boldsymbol{\alpha}_k + \boldsymbol{\beta}_k + \boldsymbol{\gamma}_k$). For 95% credible intervals, $\circ$ is lower bound; $\triangle$ is mean; and $+$ is upper bound. The number of genes assigned to each cluster is shown in parentheses. Vertical axes of the plots are not drawn to the same scale.

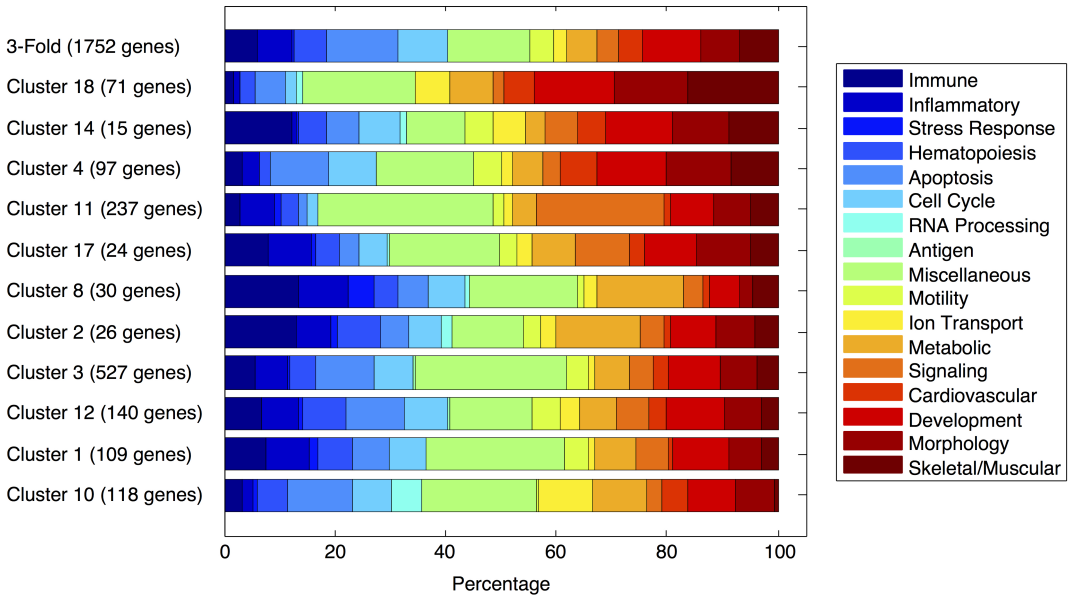


**Figure 3.** Distribution of biological functions for all genes included in the study with significant functional ontologies ($p < 0.05$); only clusters with at least 15 genes are shown; the number of genes per cluster is included in parentheses. Biological functions are based on aggregation of multiple functional ontologies with $p < 0.05$ that were considered with overlapping functions. The component functions in single categories are detailed in Supplementary Table D3. Functions of interest are placed on the left-most (e.g. immune and inflammatory-related) and the right-most (e.g. skeletal-related) sides of the scale. The middle categories represent miscellaneous functions not known to be directly related to the Brd2 mutation. The clusters are shown ordered from a low to high percentage of skeletal related functions. (This figure is reprinted with permission from the Orthopaedic Research Society [14].)

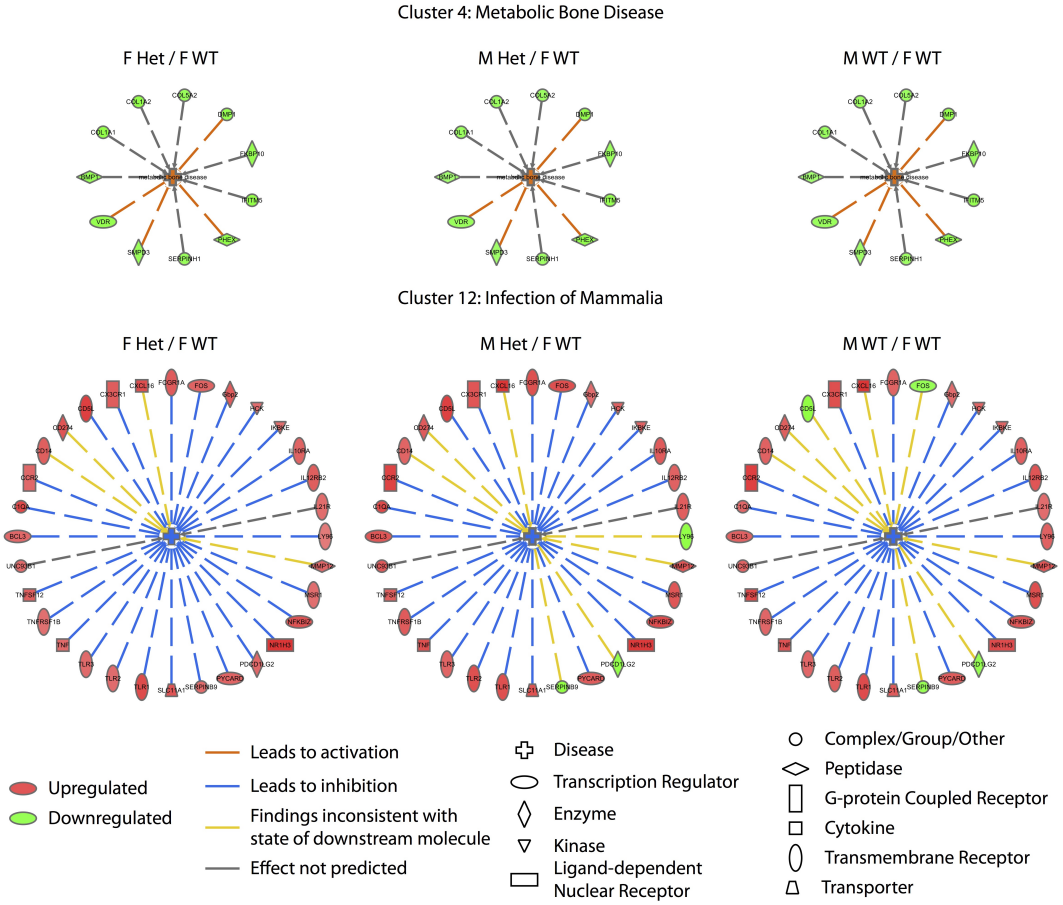


**Figure 4.** Network relationship of expressed gene groups for the highest ranked disease associated functions in clusters 4 and 12. Network interactions are as based on the IPA analysis of clusters 4 and 12. In cluster 4, the similarity of the networks in the three groups indicate that similar genetic mechanisms are turned on in the Brd2 mutant (Het) female mice, as in the males. The top-ranked functionally and biologically associated gene groups and canonical pathways are as in Supplementary Table D3.

## Appendix A. Mathematical and Computational details

### *A.1. Derivation of Reduced Likelihood Function*

First, we derive the reduced likelihood function $L(\boldsymbol{\eta}'|\boldsymbol{Y},\boldsymbol{Z})$, marginalizing over the nuisance parameters $\boldsymbol{\nu}_g$ $(g=1,\ldots,G)$. We assume that for gene $g$, $\boldsymbol{\nu}_g \sim N(\boldsymbol{m},\sigma_\nu^2\kappa_g\boldsymbol{I})$. Let us define $\boldsymbol{Y}_g=\{\boldsymbol{y}_{gijr} : i,j\in(0,1);\ r=1,\ldots,R\}$, and $\boldsymbol{Z}_g=\{z_{gk}:k=1,\ldots,K\}$, for $g=1,\ldots,G$. The part of the complete likelihood function in Eq. (2) in Section 2 relating to gene $g$ is

$$P(\boldsymbol{Y}_g,\boldsymbol{\nu}_g|\boldsymbol{\eta}',\boldsymbol{Z}_g)$$
$$\propto \prod_{k=1}^{K}\prod_{i=0}^{1}\prod_{j=0}^{1}\prod_{r=1}^{R}\left[\frac{\exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'_{gijrk}\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\varepsilon}_{gijrk}\right\}}{\sigma^T\left|\boldsymbol{\Sigma}(\rho)\right|^{1/2}}\right]^{z_{gk}}\times\frac{\exp\left\{-\frac{1}{2\sigma_\nu^2\kappa_g}(\boldsymbol{\nu}_g-\boldsymbol{m})'(\boldsymbol{\nu}_g-\boldsymbol{m})\right\}}{\sigma_\nu^T\kappa_g^{T/2}},$$

where $\boldsymbol{\varepsilon}_{gijrk}=\boldsymbol{y}_{gijr}-\boldsymbol{\nu}_g-\boldsymbol{\alpha}_k i-\boldsymbol{\beta}_k j-\boldsymbol{\gamma}_k ij$. When $z_{gk}=1$, the above can be written as,

$$\begin{aligned}P(\boldsymbol{Y}_g,\boldsymbol{\nu}_g|\boldsymbol{\eta}',z_{gk}=1) \;\propto\; & C\exp\left[-\frac{1}{2\sigma^2}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}\{\boldsymbol{\nu}_g'\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\nu}_g-2\boldsymbol{\nu}_g'\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\omega}_{gijrk}\right.\\ & \left.+\boldsymbol{\omega}'_{gijrk}\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\omega}_{gijrk}\}-\frac{1}{2\sigma_\nu^2\kappa_g}(\boldsymbol{\nu}_g-\boldsymbol{m})'(\boldsymbol{\nu}_g-\boldsymbol{m})\right],\end{aligned}$$

where $C=(\sigma^2)^{-2RT}(\sigma_\nu^2\kappa_g)^{-T/2}\left|\boldsymbol{\Sigma}(\rho)\right|^{-2R}$, $\boldsymbol{\omega}_{gijrk}=\boldsymbol{y}_{gijr}-\boldsymbol{\alpha}_k i-\boldsymbol{\beta}_k j-\boldsymbol{\gamma}_k ij$. Now, collecting similar terms in $\boldsymbol{\nu}_g$, we have

$$\begin{aligned}P(\boldsymbol{Y}_g,\boldsymbol{\nu}_g|\boldsymbol{\eta}',z_{gk}=1) \;\propto\; & C\exp\left[-\frac{1}{2\sigma_g^2}\{\boldsymbol{\nu}_g'\boldsymbol{\Lambda}_g\boldsymbol{\nu}_g-2\sigma_\nu^2\kappa_g\boldsymbol{\nu}_g'\boldsymbol{\Sigma}(\rho)^{-1}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}\boldsymbol{\omega}_{gijrk}-\right.\\ & \left.2\sigma^2\boldsymbol{\nu}_g'\boldsymbol{m}+\sigma^2\boldsymbol{m}'\boldsymbol{m}+\sigma_\nu^2\kappa_g\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}\boldsymbol{\omega}'_{gijrk}\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\omega}_{gijrk}\}\right],\end{aligned}$$

where $\sigma_g^2 = \sigma^2\sigma_\nu^2\kappa_g$ and $\boldsymbol{\Lambda}_g = 4R\sigma_\nu^2\kappa_g\boldsymbol{\Sigma}(\rho)^{-1} + \sigma^2\boldsymbol{I}$, where $\boldsymbol{\Lambda}_g$ is symmetric and non-singular. Simplifying further, this is

$$\begin{aligned}
&\propto C\exp\Bigg[-\frac{1}{2\sigma_g^2}\Bigg\{\boldsymbol{\nu}_g'\boldsymbol{\Lambda}_g\boldsymbol{\nu}_g - 2\boldsymbol{\nu}_g'\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g^{-1}\Bigg(\frac{\sigma_g^2}{\sigma^2}\boldsymbol{\Sigma}(\rho)^{-1}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}\boldsymbol{\omega}_{gijrk} \\
&\quad + \frac{\sigma_g^2}{\sigma_\nu^2\kappa_g}\boldsymbol{m}\Bigg) + \frac{\sigma_g^2}{\sigma^2}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}\boldsymbol{\omega}_{gijrk}'\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\omega}_{gijrk} + \frac{\sigma_g^2}{\sigma_\nu^2\kappa_g}\boldsymbol{m}'\boldsymbol{m}\Bigg\}\Bigg] \\
&\propto C\exp\Bigg[-\frac{1}{2\sigma_g^2}(\boldsymbol{\nu}_g - \boldsymbol{\mu}_{gk})'\boldsymbol{\Lambda}_g(\boldsymbol{\nu}_g - \boldsymbol{\mu}_{gk}) - \\
&\quad \frac{1}{2\sigma_g^2}\Bigg\{\frac{\sigma_g^2}{\sigma^2}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}\boldsymbol{\omega}_{gijrk}'\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\omega}_{gijrk} - \boldsymbol{\mu}_{gk}'\boldsymbol{\Lambda}_g\boldsymbol{\mu}_{gk} + \frac{\sigma_g^2}{\sigma_\nu^2\kappa_g}\boldsymbol{m}'\boldsymbol{m}\Bigg\}\Bigg]
\end{aligned} \tag{A1}$$

where $\boldsymbol{\mu}_{gk} = \boldsymbol{\Lambda}_g^{-1}\Bigg(\frac{\sigma_g^2}{\sigma^2}\boldsymbol{\Sigma}(\rho)^{-1}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}\boldsymbol{\omega}_{gijrk} + \sigma^2\boldsymbol{m}\Bigg)$. Now to get $P(\boldsymbol{Y}_g|\boldsymbol{\eta}', z_{gk}=1)$, we need to integrate out $\boldsymbol{\nu}_g$ from Eq.(A1). This finally gives,

$$\begin{aligned}
&P(\boldsymbol{Y}_g|\boldsymbol{\eta}', z_{gk}=1) \\
&\propto (\sigma^2)^{-\frac{T(4R-1)}{2}}|\boldsymbol{\Sigma}(\rho)|^{-2R}|\boldsymbol{\Lambda}_g|^{-1/2} \\
&\quad \times\exp\Bigg[-\frac{1}{2}\Bigg\{\frac{1}{\sigma^2}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}\boldsymbol{\omega}_{gijrk}'\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\omega}_{gijrk} - \frac{1}{\sigma_g^2}\boldsymbol{\mu}_{gk}'\boldsymbol{\Lambda}_g\boldsymbol{\mu}_{gk} + \frac{1}{\sigma_\nu^2\kappa_g}\boldsymbol{m}'\boldsymbol{m}\Bigg\}\Bigg].
\end{aligned}$$

Thus, the reduced likelihood over all $G$ genes is

$$\begin{aligned}
P(\boldsymbol{Y}|\boldsymbol{\eta}', \boldsymbol{Z}) &\propto \prod_{g=1}^{G}\prod_{k=1}^{K}P(\boldsymbol{Y}_g|\boldsymbol{\eta}', \boldsymbol{Z}_g)^{z_{gk}} \\
&\propto \prod_{g=1}^{G}\prod_{k=1}^{K}\Bigg[\Bigg(\frac{|\boldsymbol{\Sigma}(\rho)|^{-2R}|\boldsymbol{\Lambda}_g|^{-1/2}}{\sigma^{(4R-1)T}}\Bigg)^{z_{gk}} \\
&\quad \times\exp\Bigg\{-\frac{1}{2}\Bigg(\frac{1}{\sigma^2}\sum_{ijr}z_{gk}\boldsymbol{\omega}_{gijrk}'\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\omega}_{gijrk} - \frac{1}{\sigma_g^2}z_{gk}\boldsymbol{\mu}_{gk}'\boldsymbol{\Lambda}_g\boldsymbol{\mu}_{gk} + \frac{\boldsymbol{m}'\boldsymbol{m}}{\sigma_\nu^2}z_{gk}\kappa_g^{-1}\Bigg)\Bigg\}\Bigg] \\
&\propto \frac{|\boldsymbol{\Sigma}(\rho)|^{-2GR}}{\sigma^{(4R-1)GT}}\prod_{g=1}^{G}\Bigg[|\boldsymbol{\Lambda}_g|^{-1/2}\exp\Bigg\{-\frac{1}{2}\sum_{g=1}^{G}(A_g - B_g + C_g)\Bigg\}\Bigg],
\end{aligned}$$

where

$$A_g = \frac{1}{\sigma^2} \sum_{k=1}^{K} \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{r=1}^{R} \left( z_{gk} \boldsymbol{w}'_{gijrk} \boldsymbol{\Sigma}(\rho)^{-1} \boldsymbol{w}_{gijrk} \right),$$

$$B_g = \frac{1}{\sigma_g^2} \sum_{k=1}^{K} z_{gk} \boldsymbol{\mu}'_{gk} \boldsymbol{\Lambda}_g \boldsymbol{\mu}_{gk},$$

and

$$C_g = \frac{\boldsymbol{m}'\boldsymbol{m}}{\sigma_\nu^2 \kappa_g}.$$

### *A.2. Details of model-fitting through MCMC sampling*

Here we outline a Markov chain Monte Carlo (MCMC) sampling scheme for posterior estimation of the parameters of the model, consisting of both Gibbs sampling and Metropolis steps.

Step 0. Initialize parameters

$$\boldsymbol{\eta}'^{(0)} = \{\boldsymbol{\pi}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)}, \sigma^{2(0)}, \rho^{(0)}, \boldsymbol{m}^{(0)}, \sigma_\nu^{2(0)}, \sigma_\alpha^{2(0)}, \sigma_\beta^{2(0)}, \sigma_\gamma^{2(0)}\}.$$

At each iteration $s$, $(s = 1, \ldots, S)$, repeat Steps 1-11 below.

Step 1. For $k = 1, \ldots, K$, and $g = 1, \ldots, G$, evaluate the posterior cluster membership probabilities $\tilde{\pi}_{gk}^{(s)} = P(z_{gk} = 1 | \boldsymbol{\eta}'^{(s-1)}, \boldsymbol{Y})$, $(g = 1, \ldots, G,\ k = 1, \ldots, K)$, using Bayes' Rule,

$$\tilde{\pi}_{gk}^{(s)} = \frac{P(\boldsymbol{Y}_g | \boldsymbol{\eta}'^{(s-1)}, z_{gk} = 1) \pi_k^{(s-1)}}{\sum_{k=1}^{K} P(\boldsymbol{Y}_g | \boldsymbol{\eta}'^{(s-1)}, z_{gk} = 1) \pi_k^{(s-1)}}. \tag{A2}$$

Step 2. Update $(z_{g1}^{(s)}, \ldots, z_{gK}^{(s)})$ by drawing from Multinomial$(\tilde{\pi}_{g1}^{(s)}, \ldots, \tilde{\pi}_{gK}^{(s)})$, for $g = 1, \ldots, G$.

Step 3. Update $\boldsymbol{\pi}^{(s)}$ from $P(\boldsymbol{\pi} | \boldsymbol{Y}, \boldsymbol{Z}^{(s)})$, which is Dirichlet$(\theta_1 + \sum_{g=1}^{G} z_{g1}^{(s)}, \ldots, \theta_K +$

$\sum_{g=1}^{G} z_{gK}^{(s)})$ $(g = 1, \ldots, G)$.

Step 4. Update $\boldsymbol{\alpha}^{(s)}, \boldsymbol{\beta}^{(s)}$, and $\boldsymbol{\gamma}^{(s)}$ from their posterior full conditional distributions, given $\boldsymbol{Y}, \boldsymbol{Z}^{(s)}$, using Gibbs sampling. It can be shown that, for cluster $k$ $(k = 1, \ldots, K)$,

$$\boldsymbol{\alpha}_k | \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\eta}' \sim N(\boldsymbol{V}_\alpha \boldsymbol{m}_\alpha, \boldsymbol{V}_\alpha), \qquad \text{where} \tag{A3}$$

$$\begin{aligned}
\boldsymbol{V}_\alpha^{-1} &= \frac{2Rn_k}{\sigma^2}\boldsymbol{\Sigma}(\rho)^{-1} - \frac{4R^2\sigma_\nu^2}{\sigma^2}\sum_{g:z_{gk}=1}\kappa_g\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\Sigma}(\rho)^{-1} + \frac{\boldsymbol{I}}{\sigma_\alpha^2}, \\
\boldsymbol{m}_\alpha &= \frac{\boldsymbol{\Sigma}(\rho)^{-1}}{\sigma^2}\sum_{g:z_{gk}=1}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R} i\boldsymbol{a}_{gijrk} - 2R\boldsymbol{\Sigma}(\rho)^{-1}\sum_{g:z_{gk}=1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{m} \\
&\quad -\frac{2R\sigma_\nu^2}{\sigma^2}\sum_{g:z_{gk}=1}\kappa_g\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\Sigma}(\rho)^{-1}\sum_{ijr}\boldsymbol{a}_{gijrk},
\end{aligned}$$

where $n_k = \sum_{g=1}^{G} z_{gk}$, $\boldsymbol{\Lambda}_g = 4R\sigma_\nu^2\kappa_g\boldsymbol{\Sigma}(\rho)^{-1} + \sigma^2\boldsymbol{I}$, and $\boldsymbol{a}_{gijr} = \boldsymbol{y}_{gijr} - \boldsymbol{\beta}_k j - \boldsymbol{\gamma}_k ij$. Making use of the conjugacy property, the conditional posterior distributions for $\boldsymbol{\beta}_k$'s, and $\boldsymbol{\gamma}_k$'s can similarly be derived. Mathematical details of these derivations can be found in the next section.

Steps 5-11. Use Metropolis updates, separately, for $\sigma^{2(s)}, \rho^{(s)}, \sigma_\nu^{2(s)}, \sigma_\alpha^{2(s)}, \sigma_\beta^{2(s)}, \sigma_\gamma^{2(s)}, \boldsymbol{m}^{(s)}$, conditional on $\boldsymbol{Y}, \boldsymbol{Z}^{(s)}$ and all other parameters.

In pilot runs, it was found that setting the variance parameters, $(\sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2)$ to fixed large values, instead of updating them, drastically speeded up MCMC convergence without significantly affecting the parameter estimates. We also tested the performance of the algorithm under various settings of the noise variance, $\sigma_\nu^2$ (described in Section 3.1), which showed that variation in the value set for $\sigma_\nu^2$ had a negligible impact on posterior inference. Hence, unless there is strong prior information that motivates the use of informative priors for these parameters, we will use a more parsimonious Metropolis-Gibbs sampler with Steps 5-11 being replaced with the below.

Steps 5-7. Update $\sigma^{2(s)}, \rho^{(s)}, \boldsymbol{m}^{(s)}$, separately, conditional on $\boldsymbol{Y}, \boldsymbol{Z}^{(s)}$ and all other

parameters. Since closed analytical forms for the conditional posterior densities of these are not available, we use separate Metropolis steps to simulate from their posterior densities. A $\chi^2_1$ proposal density is used for $\sigma^2$, and Gaussian proposals for $\boldsymbol{m}$, and $\rho$ (after a logit transformation), tuning the parameters of the proposal densities to get approximate acceptance rates between 20% and 30%.

The steps are repeated, in turn, for iterations $s = 1, \ldots, S$, stopping when standard MCMC convergence diagnostics indicate it is safe to do so. Then we perform posterior inference on the model parameters and gene cluster memberships using the generated posterior simulations.

### *A.3. Posterior full conditional densities for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.*

Next, we illustrate the derivation of the posterior conditional distribution of $\boldsymbol{\alpha}_k$ making use of conjugacy properties. Let $\boldsymbol{\eta}'_\alpha = \boldsymbol{\eta}' \setminus \boldsymbol{\alpha} = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \sigma^2, \boldsymbol{m}, \sigma^2_\nu, \sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma)$. We have,

$$\begin{aligned} P(\boldsymbol{\alpha}_k|\boldsymbol{Y}, \boldsymbol{\eta}'_\alpha, \boldsymbol{Z}) &\propto \prod_{g:z_{gk}=1} P(\boldsymbol{Y}_g|\boldsymbol{\eta}', \boldsymbol{Z}_g) P(\boldsymbol{\alpha}_k) \\ &\propto \prod_{g:z_{gk}=1} \exp\left[ -\frac{1}{2}\left\{ \frac{1}{\sigma^2} \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R} \boldsymbol{\omega}'_{gijrk} \boldsymbol{\Sigma}(\rho)^{-1} \boldsymbol{\omega}_{gijrk} - \frac{1}{\sigma^2_g} \boldsymbol{\mu}'_{gk} \boldsymbol{\Lambda}_g \boldsymbol{\mu}_{gk} \right\}\right] \\ &\quad \times \exp\left[ -\frac{1}{2\sigma^2_\alpha} \boldsymbol{\alpha}'_k \boldsymbol{\alpha}_k \right]. \end{aligned}$$

Now, we know that $\boldsymbol{\omega}_{gijrk} = \boldsymbol{y}_{gijr} - \boldsymbol{\alpha}_k i - \boldsymbol{\beta}_k j - \boldsymbol{\gamma}_k ij$, which can be written as, $-(i\boldsymbol{\alpha}_k - \boldsymbol{a}_{gijrk})$, where $\boldsymbol{a}_{gijrk} = \boldsymbol{y}_{gijr} - \boldsymbol{\beta}_k j - \boldsymbol{\gamma}_k ij$ (i.e. the terms independent of $\boldsymbol{\alpha}_k$). This gives a new expression for $\boldsymbol{\mu}_{gk}$ as

$$\boldsymbol{\mu}_{gk} = \boldsymbol{\Lambda}_g^{-1}\left[ -\frac{\sigma^2_g}{\sigma^2} \boldsymbol{\Sigma}(\rho)^{-1} \left( 2R\boldsymbol{\alpha}_k - \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R} \boldsymbol{a}_{gijrk} \right) + \sigma^2 \boldsymbol{m} \right]. \tag{A4}$$

So now,

$$
\begin{aligned}
&P(\boldsymbol{\alpha}_k|\boldsymbol{Y},\boldsymbol{\eta}'_\alpha,\boldsymbol{Z}) \\
&\propto \exp\left[-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{g:z_{gk}=1}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}(i\boldsymbol{\alpha}_k-\boldsymbol{a}_{gijrk})'\boldsymbol{\Sigma}(\rho)^{-1}(i\boldsymbol{\alpha}_k-\boldsymbol{a}_{gijrk})\right.\right. \\
&\quad -\sum_{g:z_{gk}=1}\frac{1}{\sigma_g^2}\left\{-\frac{\sigma_g^2}{\sigma^2}\boldsymbol{\Sigma}(\rho)^{-1}(2R\boldsymbol{\alpha}_k-\sum_{ijr}\boldsymbol{a}_{gijrk})+\sigma^2\boldsymbol{m}\right\}'\boldsymbol{\Lambda}_g^{-1}\left\{-\frac{\sigma_g^2}{\sigma^2}\boldsymbol{\Sigma}(\rho)^{-1}(2R\boldsymbol{\alpha}_k-\right. \\
&\quad \left.\left.\left.\sum_{ijr}\boldsymbol{a}_{gijrk})+\sigma^2\boldsymbol{m}\right\}+\frac{\boldsymbol{\alpha}'_k\boldsymbol{\alpha}_k}{\sigma_\alpha^2}\right]\right].
\end{aligned}
\tag{A5}
$$

Now, the first term in (A5) simplifies to (ignoring terms constant with regards to $\boldsymbol{\alpha}_k$)

$$
\frac{2R}{\sigma^2}\sum_{g:z_{gk}=1}\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\alpha}_k-\frac{2}{\sigma^2}\sum_{g:z_{gk}=1}\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}i\boldsymbol{a}_{gijrk}.
$$

The second term in (A5) simplifies to

$$
\begin{aligned}
&-\sum_{g:z_{gk}=1}\left[\frac{\sigma_g^2}{\sigma^4}\left\{4R^2\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\alpha}_k-4R\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\Sigma}(\rho)^{-1}\sum_{ijr}\boldsymbol{a}_{gijrk}\right\}\right. \\
&\left.-4R\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{m}\right].
\end{aligned}
$$

Collecting similar terms in $\boldsymbol{\alpha}_k$ in (A5), we now have

$$
\begin{aligned}
&P(\boldsymbol{\alpha}_k|\boldsymbol{Y},\boldsymbol{\eta}'_\alpha,\boldsymbol{Z}) \\
&\propto \exp\left[-\frac{1}{2}\left[\frac{2R}{\sigma^2}\sum_{g:z_{gk}=1}\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\alpha}_k-\frac{4R^2}{\sigma^4}\sum_{g:z_{gk}=1}\sigma_g^2\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\alpha}_k+\frac{\boldsymbol{\alpha}'_k\boldsymbol{\alpha}_k}{\sigma_\alpha^2}\right.\right. \\
&\quad -2\left\{\frac{1}{\sigma^2}\sum_{g:z_{gk}=1}\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R}i\boldsymbol{a}_{gijrk}-\frac{2R}{\sigma^4}\sum_{g:z_{gk}=1}\sigma_g^2\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\Sigma}(\rho)^{-1}\sum_{ijr}\boldsymbol{a}_{gijrk}\right\} \\
&\quad \left.\left.+4R\boldsymbol{\alpha}'_k\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{m}\right]\right].
\end{aligned}
\tag{A6}
$$

Now, set $n_k = \sum_{g=1}^{G} z_{gk}$, the number of genes in cluster $k$ ($k = 1, \ldots, K$). Then (A6) gives, after further simplification,

$$P(\boldsymbol{\alpha}_k|\boldsymbol{Y}, \boldsymbol{\eta}'_\alpha, \boldsymbol{Z}) \quad \propto \quad \exp\left[ -\frac{1}{2}\left\{ \boldsymbol{\alpha}'_k \boldsymbol{V}_\alpha^{-1} \boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}'_k \boldsymbol{m}_\alpha \right\}\right],$$

where

$$\begin{aligned}
\boldsymbol{V}_\alpha^{-1} &= \frac{2Rn_k}{\sigma^2}\boldsymbol{\Sigma}(\rho)^{-1} - \frac{4R^2\sigma_\nu^2}{\sigma^2}\sum_{g:z_{gk}=1}\kappa_g \boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\Sigma}(\rho)^{-1} + \frac{\boldsymbol{I}}{\sigma_\alpha^2}, \\
\boldsymbol{m}_\alpha &= \frac{\boldsymbol{\Sigma}(\rho)^{-1}}{\sigma^2}\sum_{g:z_{gk}=1}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=1}^{R} i\boldsymbol{a}_{gijrk} - \frac{2R\sigma_\nu^2}{\sigma^2}\sum_{g:z_{gk}=1}\kappa_g\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\Sigma}(\rho)^{-1}\sum_{ijr}\boldsymbol{a}_{gijrk} \\
&\quad -2R\boldsymbol{\Sigma}(\rho)^{-1}\sum_{g:z_{gk}=1}\boldsymbol{\Lambda}_g^{-1}\boldsymbol{m},
\end{aligned}$$

where, as previously, $\boldsymbol{\Lambda}_g = 4R\sigma_\nu^2\kappa_g\boldsymbol{\Sigma}(\rho)^{-1} + \sigma^2\boldsymbol{I}$, for $g = 1, \ldots, G$. So, it can be inferred that

$$\boldsymbol{\alpha}_k|\boldsymbol{Y}, \boldsymbol{\eta}'_\alpha, \boldsymbol{Z} \sim N(\boldsymbol{V}_\alpha\boldsymbol{m}_\alpha, \boldsymbol{V}_\alpha).$$

Similar derivations for the posterior full conditional distributions are carried out for both $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$.

### *A.4. Details of sampling procedures for other parameters.*

The remaining parameters to be estimated through MCMC are $\{\boldsymbol{m}, \rho, \sigma^2\}$. $\boldsymbol{m}$ can be updated using a Metropolis algorithm, where at iteration $(s+1)$, we choose a symmetric proposal density $N(\boldsymbol{m}^{(s)}, \zeta\boldsymbol{I})$, where $\boldsymbol{m}^{(s)}$ is the sampled value of $\boldsymbol{m}$ at iteration $t$, and $\zeta$ is chosen through pilot runs to give an acceptance rate of between 20-50%. For updating $\sigma^2$, a $\chi_1^2$ distribution is chosen as a proposal density. For $\rho$, the selection of proposal densities is limited by the necessity that $-1 \leq \rho \leq 1$. One possibility is to use a Uniform$(-1, 1)$ proposal density, another, is to transform $\rho$, so

that it takes values on the real line, using the logit transformation,

$$\psi = \log\left(\frac{1+\rho}{1-\rho}\right).$$

We then update $\psi$ using the Metropolis algorithm with the target density

$$P(\psi|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\eta'}) \propto P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\eta'}(\psi))J(\rho;\psi)P(\psi)$$

where $P(\psi)$ is a non-informative (improper uniform) prior density on $\psi$ and $\boldsymbol{\eta'}(\psi)$ denotes all parameters in $\boldsymbol{\eta'}$ while replacing $\rho$ by $\psi$. At the $(s+1)^{\text{th}}$ step of the sampler, we use a Normal proposal density for $\psi$, centered at $\psi^{(s)}$ and with a variance tuned to provide a reasonable acceptance rate. In practice it is observed that a logit transformation of $\rho$ leads to better convergence and a more efficient sampler than directly sampling $\rho$.